SCHOOL OF MEDICINE UNIVERSITY OF THESSALY

POSTGRADUATE PROGRAMME (MSC)

«RESEARCH METHODOLOGY IN BIOMEDICINE, BIOSTATISTICS AND CLINICAL BIOINFORMATICS AT UNIVERSITY OF THESSALY»

Master's Thesis

"Multicollinearity: diagnostics and ridge regression as a method of handling"

"Πολυσυγγραμμικότητα: διαγνωστικές μέθοδοι και η ridge regression ως μέθοδος αντιμετώπισης"

Supervisors

Batsidis Apostolos, Assistant Professor

Stefanidis Ioannis, Professor

Doxani Chrysoula, Research Fellow

Kanellopoulou Aikaterini

AEM: M060616013

E-mail: katerkane@gmail.com

Academic year: 2016 – 2017

Contents

ABSTRACT

The problem of multicollinearity occurs when two or more independent variables in a regression model are highly correlated. The main consequence of multicollinearity is that the parameter estimates are less precise. In this dissertation, we referred to the multicollinearity problem, the methods of identifying this problem and the ridge regression as a method of handling multicollinearity. A real data set from biochemistry was used to illustrate how the method is applied.

KEYWORDS: Multicollinearity, Multiple linear regression, Ridge regression

SECTION 1: INTRODUCTION

Multiple linear regression is a statistical technique that allows researchers to predict someone's score on one variable taking into account their scores on several other variables (Saleh, 2014). For example, in order to estimate the body fat (dependent or response variable), we can utilize three independent variables or predictors, i.e. the triceps skinfold thickness, the thigh circumference, and the midarm circumference. Therefore, in multiple regression we attempt to predict a dependent or response variable based on an assumed linear relation with several independent or predictor variables.

The multiple linear regression model is expressed using matrix notation in the following form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of responses, $\boldsymbol{X}$ is an $n \times p$ matrix of the independent variables (predictors), $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors. The estimates of $\boldsymbol{\beta}$ coefficients are, based on the least square method, the values that minimize the sum of squared errors for the sample. It is easily proved after some algebra that

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

is the least square estimate of $\boldsymbol{\beta}$ where $\boldsymbol{X}'$ is the transpose of the matrix $\boldsymbol{X}$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is the inverse of the $\boldsymbol{X}'\boldsymbol{X}$.

A problem that can be emerged when we fit a multiple linear regression model is multicollinearity. This is a statistical phenomenon, in which a perfect or exact relationship between the independent variables exists. Thus, *multicollinearity* (also known as *collinearity*) is a phenomenon in which one predictor variables in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy (https://en.wikipedia.org/wiki/Multicollinearity). When there is a perfect or exact relationship between the independent variables, it is difficult to come up with reliable estimates of their individual coefficients. It will result in incorrect conclusions

about the relationship between outcome variable and independent variables. In such a case, the matrix $\boldsymbol{X'X}$ cannot be inverted.

Remark There are two types of multicollinearity (https://onlinecourses.science.psu.edu/stat501/node/344): structural multicollinearity, which is a mathematical artifact caused by creating new predictors from other predictors and data-based multicollinearity, which is a result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected. This is the type we encounter most often!

One consequence of a high degree of multicollinearity is that, even if the matrix $\boldsymbol{X'X}$ is invertible, a computer algorithm may be unsuccessful in obtaining an approximate inverse, and if it does obtain one it may be numerically inaccurate. But even in the presence of an accurate $\boldsymbol{X'X}$ matrix, the presence of multicollinearity can cause serious problems with the estimation of $\boldsymbol{\beta}$ and the interpretation. Multicollinearity inflates the variances of the parameter estimates and hence this may lead to lack of statistical significance of individual independent variables even though the overall model may be significant. The greater the multicollinearity, the greater the standard errors. When high multicollinearity is present, confidence intervals for coefficients tend to be very wide and t-statistics tend to be very small. Coefficients will have to be larger in order to be statistically significant, i.e. it will be harder to reject the null when multicollinearity is present. However, things besides multicollinearity can cause large standard errors. When two independent variables are highly and positively correlated, their slope coefficient estimators will tend to be highly and negatively correlated. Related to the interpretation, the usual interpretation of a regression coefficient is that it provides an estimate of the effect of a one unit change in an independent variable, say X_1, holding the other variables constant. If X_1 is highly correlated with another independent variable, say X_2, in the given data set, then we have a set of observations for which X_1 and X_2 have a particular linear stochastic relationship. We do not have a set of observations for which all changes in X_1 are independent of changes in X_2, so we have an imprecise estimate of the effect of independent changes in X_1.

In this dissertation, we will first describe methods to diagnose the problem of multicollinearity and then we will describe ridge regression as a method of handling multicollinearity.

SECTION 2: METHODS TO IDENTIFY MULTICOLLINEARITY

Indicators that multicollinearity may be present in a model include: (i) large changes in the estimated regression coefficients when a predictor variable is added or deleted and (ii) insignificant regression coefficients for the affected variables in the multiple regression, but a rejection, using an F-test, of the joint hypothesis that those coefficients are all zero.

On the other hand, there are three basic methods to identify multicollinearity: examination of correlation matrix, Variance Inflation Factor (VIF), and eigensystem analysis of correlation matrix. In the sequel, we will briefly describe these three methods of multicollinearity identification.

Examination of Correlation Matrix

The correlation matrix of n random variables $X_1, \ldots, X_n$ is the $n \times n$ matrix whose i, j entry is the correlation of X_i, X_j. The correlation matrix is symmetric because the correlation between X_i and X_j is the same as the correlation between X_j and X_i.

Large correlation coefficients in the correlation matrix of independent variables indicate multicollinearity. If there is a multicollinearity between any two independent variables, then the absolute value of the correlation coefficient between these two variables will be near to unity. Thus, multicollinearity is identified whenever large correlation coefficients are found in the correlation matrix of independent variables. For example, consider the correlation matrix of Table 1. We observe that the variables X_1 and X_2 are highly correlated (Pearson's Correlation coefficient= 0.935). This is a sign that multicollinearity is present. Statistical inference based on Pearson's correlation coefficient focuses testing that there is no linear correlation between the two variable. In this context in Table 1 the p-values of the 2-tailed hypothesis are given.

Thus, when pairwise correlations exceed a threshold or is statistical significant based on the previous mentioned hypothesis test, collinearity is considered high. Dormann et

al. (2012) considered values greater than 0.7 as an appropriate indicator that collinearity begins to severely distort model estimation and subsequent prediction.

Remark The procedure based on the correlation matrix is often highly problematic since correlation describes a bivariate relationship, whereas collinearity is a multivariate phenomenon.

Correlations

		Y	X1	X2	X3
Pearson Correlation		1	-.479**	-.490**	.083
Sig. (2-tailed)			.000	.000	.410
N		100	100	100	100
Pearson Correlation		-.479**	1	.935**	-.062
Sig. (2-tailed)		.000		.000	.542
N		100	100	100	100
Pearson Correlation		-.490**	.935**	1	-.036
Sig. (2-tailed)		.000	.000		.719
N		100	100	100	100
Pearson Correlation		.083	-.062	-.036	1
Sig. (2-tailed)		.410	.542	.719	
N		100	100	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Table 1

Variance Inflation Factor

Another method to quantify the severity of multicollinearity in an ordinary least square regression analysis is the Variance Inflation Factor (*VIF*). Let R_j^2 denote the determination coefficient (i.e. the proportion of the variance in the dependent variable that is predictable from the independent variables) when the independent variable X_j is regressed on all other independent variables in the model. The *VIF* index is defined as

$$VIF_j = 1/(1 - R_j^2),$$

for $j = 1,2, \dots, p - 1$ (Seber and Lee, 2003).

It is obvious that VIF_j equals to 1 when the *j*-th variable is not linearly related to the other independent variables, i.e. when R_j^2 equals to 0. On the other hand, the VIF_j tends to infinity when the *j*-th variable is linearly related to the other independent variables, i.e. when R_j^2 tends to 1. Thus, the *VIF* index measures how much variance of an estimated regression coefficient is increased because of multicollinearity.

A rule of thumb is that if any of the *VIF* values exceeds 5 or 10, that the associated regression coefficients are poorly estimated because of multicollinearity (Montgomery et al., 2001). For example, consider the results of a regression model fitting shown in Table 2. We observe that the VIF index for variable X1 equals to 8.041 while for variable X2 equals to 8.021. These high values of VIF indicate the presence of multicollinearity. However, according to O'Brien (2007), the rules of thumb for *VIF* indexes should be used with caution.

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	.543	.051		10.617	.000		
X1	-.159	.262	-.152	-.607	.545	.124	8.041
X2	-.373	.271	-.345	-1.374	.173	.125	8.021
X3	.054	.079	.061	.689	.492	.993	1.007

a. Dependent Variable: Y

Table 2

A similar measure is *tolerance* defined as

$$\frac{1}{VIF}.$$

According to Dormann et al. (2012), values of tolerance smaller than 0.1 are sign for multicollinearity. The tolerance values are also shown in Table 2.

Eigensystem Analysis of Correlation Matrix

Several other multicollinearity diagnostics are related with the eigenvalues of the correlation matrix (Dormann et al., 2012). Assume that we have a $p \times p$ matrix $\boldsymbol{A}$. The p *eigenvalues* of $\boldsymbol{A}$, denoted as $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the solution of the equation

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0,$$

where $\boldsymbol{I}$ is the identity matrix.

If any one of the eigenvalues is exactly equal to zero, there is a perfect linear relationship among the original variables, which is an extreme case of multicollinearity (Chatterjee and Hadi, 2006). Moreover, if the *determinant of correlation matrix* (*D*), which equals the product of the eigenvalues, is close to zero collinearity is high.

Another measure of the overall multicollinearity of the variables can be obtained by computing the *condition number* (*CN*) of the correlation matrix, defined by

$$\frac{maximum\ eigenvalue\ of\ the\ correlation\ matrix}{minimum\ eigenvalue\ of\ the\ correlation\ matrix} = \frac{\lambda_{max}}{\lambda_{min}}$$

(Chatterjee and Hadi, 2006). The condition number will always be greater than 1. A large condition number indicates evidence of strong collinearity. The harmful effects of collinearity in the data become strong when the values of the condition number exceeds 15 while corrective action should always be taken when the *CN* exceeds 30.

Moreover, if multicollinearity is present in the independent variables, one or more of the eigenvalues will be small (near to zero). A rule of thumb is that if one or more of the eigenvalues are small (close to zero) and the corresponding condition number is large, then there is multicollinearity (Montgomery et al., 2001). As an example, consider the results shown in Table 3. The eigenvalues of this matrix are 2.825, 0.976, 0.151, and 0.047. We observe that the last eigenvalue is close to zero which means that multicollinearity is possible present

Collinearity Diagnostics[a]

			Variance Proportions			
			(Constant)	X1	X2	X3
1	2,825	1,000	,02	,01	,01	,02
2	,976	1,701	,05	,02	,02	,10
3	,151	4,321	,91	,00	,01	,87
4	,047	7,739	,01	,97	,96	,01

a. Dependent Variable: Y

Table 3

In the previous context, the *condition index* (CI) is a measure of severity of multicollinearity associated with j-th eigenvalues (Belsley et al., 1980; Johnston, 1984; Douglass et al., 2003). The CIs of a correlation matrix are the square roots of the ratios of the largest eigenvalue divided by the one in focus. CIs equal or larger than 30 are considered "large" and critical (Dormann et al., 2012).

The *variance-decomposition proportions* (*VD*) are the variance proportions of the i-th variable attributable to the j-th eigenvalue (Booth et al., 1994; Belsley, 1991). No variable should attribute more than 0.5 to any one eigenvalue (Dormann et al., 2012).

The square root of CN is another measure for multicollinearity. It is denoted by

$$K = \sqrt{\frac{maximum\ eigenvalue\ of\ the\ correlation\ matrix}{minimum\ eigenvalue\ of\ the\ correlation\ matrix}} = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

and multicollinearity is present if K is greater than 5. For the example of Table 3, it holds that

$$K = \sqrt{\frac{2.825}{0.047}} = 7.753,$$

which implies multicollinearity.

Another empirical criterion for the presence of multicollinearity is given by the sum of the reciprocals of the eigenvalues, that is,

$$\sum_{j=1}^{p} \frac{1}{\lambda_j}.$$

If the sum is greater than, say, five times the number of independent variables, multicollinearity is present (Chatterjee and Hadi, 2006). For the example of Table 3, we have

$$\sum_{j=1}^{p} \frac{1}{\lambda_j} = \frac{1}{2.825} + \frac{1}{0.976} + \frac{1}{0.151} + \frac{1}{0.047} = 29.278.$$

Given that the above sum is greater than five times the number of independent variables, i.e. $3 \times 5 = 15$, we assume that multicollinearity is present.

SECTION 3: RIDGE REGRESSION

As we already mentioned in the previous section, the presence of multicollinearity in least squares regression cause larger variances of parameter estimates with a consequence the parameter estimates to be less precise (Adnan et al., 2006). Therefore, the more the multicollinearity, the less interpretable are the parameters.

In order to diminish multicollinearity, we have to exclude one or more independent variables from the multiple regression model. However, this is not always possible. In the case where none of the independent variables can be dropped, there are two alternative methods: ridge regression and principal component regression. Adnan et al. (2006) compared the performances of ridge regression, principal component regression and partial least squares regression for handling the multicollinearity problem. In this dissertation we will deal only with ridge regression.

Ridge regression is a popular parameter estimation method used to address the collinearity problem frequently arising in multiple linear regression (McDonald, 2009). Hoerl and Kennard (1970), as an alternative procedure to the OLS method in regression analysis, suggested ridge regression, especially, when multicollinearity exists (Saleh, 2014). In other words, ridge regression is a modification of the least squares method proposed by that allows biased estimators of the regression coefficients. Although the estimators are biased, the biases are small enough for these estimators to be substantially more precise than unbiased estimators are (Adnan et al., 2006). Therefore, these biased estimators are preferred over unbiased ones since they will have a larger probability of being close to the true parameter values.

Hoerl and Kennard (1970) showed that the estimates of regression coefficients tend to become too large in absolute values, and it is possible that some will even have the wrong sign. The chances of encountering such difficulties increase the more the prediction vectors deviate from orthogonality.

Recall the standard multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon_i,$$

or using matrix notation in the following form

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}.$$

Let

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

be the least square estimate of $\boldsymbol{\beta}$. The term "least squares" means the estimation of $\boldsymbol{\beta}$ is based on the minimization of the sum of the squares of the residuals ε_i. The prime "′" on $\boldsymbol{X'}$ is an operator which flips matrix $\boldsymbol{X}$ over its diagonal, i.e. it switches the row and column indices of the matrix, producing a new matrix.

In ridge regression, the first step is to standardize both the dependent variable and the independent variables by subtracting their means and dividing by their standard deviations. In order not to use too many notation, in the sequel we will assume that $\boldsymbol{Y}$ and $\boldsymbol{X}$ are the standardized dependent and independent variables respectively.

As we already mentioned, the least square estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}.$$

Because the variables are standardized, it holds that

$$\boldsymbol{X'X} = \boldsymbol{R},$$

where $\boldsymbol{R}$ is the correlation matrix of the independent variables. The estimates $\widehat{\boldsymbol{\beta}}$ are unbiased, which means that the expected value of the estimates are the population values, i.e.

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

The variance-covariance matrix of $\boldsymbol{\beta}$ is

$$V(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X'X})^{-1} = \boldsymbol{R^{-1}}.$$

From the above equation, we have that

$$V(\beta_j) = 1/(1 - R_j^2),$$

which is the *VIF* index presented in Section 2.

The rational of ridge regression is to add a nonnegative constant k to matrix $\boldsymbol{X'X} = \boldsymbol{R}$. In other words, ridge regression is an estimation procedure based upon

$$\widehat{\boldsymbol{\beta}}(k) = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'Y} = (\boldsymbol{R} + k\boldsymbol{I})^{-1}\boldsymbol{X'Y}.$$

In simple words, ridge regression adds a non-negative value k, usually less than 0.3, to the diagonal elements of the correlation matrix. The method got its name by the fact that the diagonal of ones in the correlation matrix may be thought of as a ridge (a long narrow hilltop).

The amount of bias in the ridge estimator $\widehat{\boldsymbol{\beta}}(k)$ is given by

$$E\left(\widehat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta}(k)\right) = [(\boldsymbol{R} + k\boldsymbol{I})^{-1}\boldsymbol{R} - \boldsymbol{I}]\boldsymbol{\beta}(k)$$

while the covariance matrix is

$$V\left(\widehat{\boldsymbol{\beta}}(k)\right) = (\boldsymbol{R} + k\boldsymbol{I})^{-1}\boldsymbol{R}(\boldsymbol{R} + k\boldsymbol{I})^{-1}.$$

This means that the ridge estimators $\widehat{\boldsymbol{\beta}}(k)$ are biased but tend to have a smaller mean squared error than least estimators $\widehat{\boldsymbol{\beta}}$ estimators (Hoerl and Kennard, 1970).

It can be shown that there exists a value of k for which the mean squared error (i.e. the variance plus the bias squared) of the ridge estimator $\widehat{\boldsymbol{\beta}}(k)$ is less than that of the least squares estimator.

According to Chatterjee and Hadi (2006), the idea of ridge regression is to pick a value of k for which the reduction in total variance is not exceeded by the increase in bias. Technical details on ridge regression can be found on Chatterjee and Hadi (2006). For the properties of the ridge estimator, the interested reader is referred to McDonald (2009).

In the next section we will present an application of ridge regression on a real data set in order to illustrate its use and utility.

SECTION 4: APPLICATION OF RIDGE REGRESSION IN REAL DATA

In this section, we will use a real data set from biochemistry in order to illustrate the use and utility of the ridge regression method. Several biochemical indices were recorded for 163 students, 4-17 years old, from Sparta, Laconia and Leonidio, Arcadia. (Maggana, 2016). The purpose of this study was to investigate the relationship between childhood obesity and endothelial dysfunction as well as the interplay among the two pathological entities and apoptosis.

The main objective is to fit a regression model in order to estimate cholesterol (CHOL) using as predictors C-reactive protein (CRP), high-density lipoprotein (HDL), potassium (K[1]), triglycerides (TRG), phosphorus (P), chromium (CR), high-density lipoprotein (LDL), mean volume of platelets (MPV), and the cholesterol ratio CAD.

We will first fit a multiple linear regression model with cholesterol (CHOL) as dependent variable and CRP, HDL, KA, TRG, P, CR, LDL, MPV, CAD as predictors. The model is a significant model ($F(9,131) = 633.761, p < 0.001$) with adjusted R^2 equal to 0.976.

Table 4 shows the regression coefficients along with the collinearity statistics (tolerance and VIF values). We observe that all predictors are statistically significant (i.e. $p \leq 0.035$). Thus the model with which we can estimate the total cholesterol is

$$CHOL = 24.688 + 0.389HDL + 1.344LDL + 0.070TRG - 8.059CR - 3.145P + 5.001KA + 0.567MPV - 8.375CAD - 4.462CRP.$$

However, this model may not be useful as the high value of VIF for CAD (equals 8.579) indicates the presence of multicollinearity. This is rational as CAD is the CHOL/LDL ratio. Moderate values of the VIF index we also have for HDL and LDL (3.847 and 3.770, respectively).

[1] Potassium is normally denoted as K. However, in this dissertation we will use the notation KA in order to avoid confusion with the parameter K of ridge regression.

Moreover, examining the eigenvalues presented in Table 5, we observe that the last six eigenvalues are close to zero. This is also a sign that multicollinearity is possible present.

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	24,668	7,120		3,465	,001		
HDL	,389	,052	,193	7,515	,000	,260	3,847
LDL	1,344	,034	1,001	39,401	,000	,265	3,770
TRG	,070	,012	,106	5,782	,000	,510	1,961
CR	-8,059	3,781	-,033	-2,131	,035	,711	1,407
P	-3,145	,858	-,054	-3,666	,000	,802	1,247
KA	5,001	,967	,071	5,172	,000	,900	1,111
MPV	,567	,201	,050	2,823	,005	,555	1,801
CAD	-8,375	1,594	-,201	-5,255	,000	,117	8,579
CRP	-4,462	1,958	-,035	-2,279	,024	,730	1,371

a. Dependent Variable: CHOL

Table 4

Collinearity Diagnostics[a]

			Variance Proportions									
			(Constant)	HDL	LDL	TRG	CR	P	KA	MPV	CAD	CRP
1	9,215	1,000	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
2	,408	4,754	,00	,00	,00	,03	,00	,00	,00	,01	,00	,50
3	,209	6,642	,00	,00	,00	,43	,00	,00	,00	,01	,00	,10
4	,082	10,616	,00	,05	,00	,07	,00	,00	,00	,24	,01	,14
5	,037	15,676	,00	,02	,09	,23	,02	,00	,00	,26	,05	,01
6	,021	20,808	,01	,08	,11	,00	,24	,01	,02	,36	,00	,10
7	,016	24,186	,00	,00	,10	,04	,50	,12	,03	,03	,00	,09
8	,006	39,759	,00	,03	,02	,00	,08	,61	,54	,02	,02	,01
9	,004	45,705	,07	,34	,36	,02	,00	,10	,36	,08	,40	,00
10	,002	72,051	,92	,48	,32	,18	,15	,15	,05	,00	,51	,05

a. Dependent Variable: CHOL

Table 5

Thus, we will proceed with a ridge regression analysis. We run an appropriate SPSS syntax.

We first present a correlation matrix (Table 6). Highlighted are all the statistically significant correlations. CAD is highly correlated with HDL ($r = -0.623, p < 0.001$), LDL ($r = 0.651, p < 0.001$), and TRG ($r = 0.610, p < 0.001$). A moderate correlation also holds between TRG and LDL ($r = 0.495, p < 0.001$) and between MPV and CR ($r = 0.411, p < 0.001$).

Correlations[a]

	CHOL	HDL	LDL	TRG	CR	P	KA	MPV	CAD	CRP
Pearson Correlation	1	,297	,942	,446	,178	,153	,203	,125	,429	-,262
Sig. (2-tailed)		,000	,000	,000	,035	,071	,016	,140	,000	,002
Pearson Correlation	,297	1	,022	-,200	-,143	,243	,071	-,359	-,623	,031
Sig. (2-tailed)	,000		,800	,018	,091	,004	,401	,000	,000	,711
Pearson Correlation	,942	,022	1	,495	,212	,129	,140	,206	,651	-,262
Sig. (2-tailed)	,000	,800		,000	,012	,127	,098	,014	,000	,002
Pearson Correlation	,446	-,200	,495	1	,296	,103	,116	,187	,610	-,098
Sig. (2-tailed)	,000	,018	,000		,000	,225	,170	,026	,000	,246
Pearson Correlation	,178	-,143	,212	,296	1	-,205	,084	,411	,267	-,312
Sig. (2-tailed)	,035	,091	,012	,000		,015	,323	,000	,001	,000
Pearson Correlation	,153	,243	,129	,103	-,205	1	,227	-,258	-,063	,112
Sig. (2-tailed)	,071	,004	,127	,225	,015		,007	,002	,456	,186
Pearson Correlation	,203	,071	,140	,116	,084	,227	1	,006	,091	,060
Sig. (2-tailed)	,016	,401	,098	,170	,323	,007		,945	,283	,481
Pearson Correlation	,125	-,359	,206	,187	,411	-,258	,006	1	,483	-,430
Sig. (2-tailed)	,140	,000	,014	,026	,000	,002	,945		,000	,000
Pearson Correlation	,429	-,623	,651	,610	,267	-,063	,091	,483	1	-,260
Sig. (2-tailed)	,000	,000	,000	,000	,001	,456	,283	,000		,002
Pearson Correlation	-,262	,031	-,262	-,098	-,312	,112	,060	-,430	-,260	1
Sig. (2-tailed)	,002	,711	,002	,246	,000	,186	,481	,000	,002	

a. Listwise N=141

Table 6

Tables 7a – 7c show the R^2 value and the beta coefficients for estimated values of the parameter k.

R-SQUARE AND BETA COEFFICIENTS FOR ESTIMATED VALUES OF K										
K	RSQ	HDL	LDL	TRG	CR	P	KA	MPV	CAD	CRP
0.00	0.978	0.193	1.001	0.106	-0.033	-0.054	0.071	0.050	-0.201	-0.035
0.01	0.977	0.218	0.963	0.099	-0.029	-0.051	0.070	0.043	-0.153	-0.035
0.02	0.976	0.237	0.932	0.094	-0.025	-0.048	0.068	0.037	-0.116	-0.036
0.03	0.975	0.250	0.906	0.091	-0.022	-0.046	0.067	0.033	-0.086	-0.037
0.04	0.974	0.261	0.883	0.089	-0.019	-0.044	0.067	0.029	-0.062	-0.038
0.05	0.972	0.269	0.862	0.088	-0.017	-0.041	0.066	0.026	-0.043	-0.039
0.06	0.970	0.275	0.844	0.087	-0.015	-0.039	0.066	0.024	-0.026	-0.040
0.07	0.969	0.280	0.827	0.086	-0.014	-0.037	0.065	0.022	-0.012	-0.041
0.08	0.967	0.284	0.812	0.086	-0.012	-0.035	0.065	0.020	-0.001	-0.042
0.09	0.965	0.287	0.798	0.086	-0.011	-0.033	0.065	0.018	0.010	-0.043
0.10	0.963	0.289	0.785	0.086	-0.009	-0.031	0.064	0.017	0.018	-0.044
0.11	0.962	0.290	0.773	0.087	-0.008	-0.029	0.064	0.015	0.026	-0.045
0.12	0.960	0.291	0.761	0.087	-0.007	-0.027	0.064	0.014	0.033	-0.046
0.13	0.958	0.292	0.750	0.087	-0.006	-0.025	0.064	0.013	0.039	-0.047
0.14	0.956	0.292	0.740	0.088	-0.005	-0.024	0.064	0.012	0.045	-0.048
0.15	0.954	0.292	0.730	0.088	-0.004	-0.022	0.063	0.012	0.050	-0.049
0.16	0.952	0.291	0.721	0.089	-0.003	-0.020	0.063	0.011	0.054	-0.049
0.17	0.949	0.291	0.712	0.090	-0.002	-0.019	0.063	0.010	0.058	-0.050
0.18	0.947	0.290	0.703	0.090	-0.002	-0.017	0.063	0.010	0.061	-0.051
0.19	0.945	0.289	0.695	0.091	-0.001	-0.016	0.063	0.009	0.065	-0.052
0.20	0.943	0.288	0.687	0.091	0.000	-0.015	0.063	0.009	0.067	-0.052
0.21	0.941	0.287	0.680	0.092	0.000	-0.013	0.063	0.008	0.070	-0.053
0.22	0.939	0.286	0.672	0.093	0.001	-0.012	0.063	0.008	0.073	-0.054
0.23	0.937	0.285	0.665	0.093	0.002	-0.011	0.063	0.008	0.075	-0.054
0.24	0.934	0.284	0.658	0.094	0.002	-0.009	0.063	0.007	0.077	-0.055
0.25	0.932	0.282	0.651	0.094	0.003	-0.008	0.062	0.007	0.079	-0.055
0.26	0.930	0.281	0.645	0.095	0.004	-0.007	0.062	0.007	0.080	-0.056
0.27	0.928	0.279	0.638	0.095	0.004	-0.006	0.062	0.007	0.082	-0.056
0.28	0.925	0.278	0.632	0.096	0.005	-0.005	0.062	0.006	0.084	-0.057
0.29	0.923	0.276	0.626	0.096	0.005	-0.004	0.062	0.006	0.085	-0.057
0.30	0.921	0.275	0.621	0.097	0.006	-0.003	0.062	0.006	0.086	-0.058
0.31	0.918	0.273	0.615	0.097	0.006	-0.002	0.062	0.006	0.087	-0.058
0.32	0.916	0.272	0.609	0.098	0.007	-0.001	0.062	0.006	0.088	-0.059
0.33	0.914	0.270	0.604	0.098	0.007	0.000	0.062	0.006	0.089	-0.059
0.34	0.911	0.269	0.599	0.098	0.008	0.001	0.062	0.005	0.090	-0.059
0.35	0.909	0.267	0.593	0.099	0.008	0.002	0.062	0.005	0.091	-0.060

Table 7a

R-SQUARE AND BETA COEFFICIENTS FOR ESTIMATED VALUES OF K										
k	RSQ	HDL	LDL	TRG	CR	P	KA	MPV	CAD	CRP
0.36	0.907	0.265	0.588	0.099	0.008	0.003	0.062	0.005	0.092	-0.060
0.37	0.905	0.264	0.584	0.100	0.009	0.003	0.061	0.005	0.093	-0.060
0.38	0.902	0.262	0.579	0.100	0.009	0.004	0.061	0.005	0.093	-0.061
0.39	0.900	0.261	0.574	0.100	0.010	0.005	0.061	0.005	0.094	-0.061
0.40	0.898	0.259	0.569	0.100	0.010	0.006	0.061	0.005	0.094	-0.061
0.41	0.895	0.257	0.565	0.101	0.010	0.006	0.061	0.005	0.095	-0.061
0.42	0.893	0.256	0.561	0.101	0.011	0.007	0.061	0.005	0.095	-0.062
0.43	0.891	0.254	0.556	0.101	0.011	0.008	0.061	0.005	0.096	-0.062
0.44	0.888	0.253	0.552	0.102	0.011	0.009	0.061	0.005	0.096	-0.062
0.45	0.886	0.251	0.548	0.102	0.012	0.009	0.061	0.005	0.097	-0.062
0.46	0.884	0.250	0.544	0.102	0.012	0.010	0.061	0.005	0.097	-0.062
0.47	0.881	0.248	0.540	0.102	0.012	0.010	0.061	0.005	0.097	-0.062
0.48	0.879	0.247	0.536	0.102	0.013	0.011	0.060	0.005	0.098	-0.063
0.49	0.877	0.245	0.532	0.103	0.013	0.012	0.060	0.005	0.098	-0.063
0.50	0.874	0.244	0.528	0.103	0.013	0.012	0.060	0.005	0.098	-0.063
0.51	0.872	0.242	0.524	0.103	0.013	0.013	0.060	0.005	0.098	-0.063
0.52	0.870	0.241	0.521	0.103	0.014	0.013	0.060	0.005	0.099	-0.063
0.53	0.868	0.239	0.517	0.103	0.014	0.014	0.060	0.005	0.099	-0.063
0.54	0.865	0.238	0.514	0.103	0.014	0.014	0.060	0.005	0.099	-0.063
0.55	0.863	0.236	0.510	0.103	0.014	0.015	0.060	0.005	0.099	-0.063
0.56	0.861	0.235	0.507	0.103	0.015	0.015	0.060	0.005	0.099	-0.064
0.57	0.858	0.233	0.503	0.104	0.015	0.016	0.059	0.005	0.099	-0.064
0.58	0.856	0.232	0.500	0.104	0.015	0.016	0.059	0.005	0.100	-0.064
0.59	0.854	0.231	0.497	0.104	0.015	0.016	0.059	0.005	0.100	-0.064
0.60	0.852	0.229	0.494	0.104	0.016	0.017	0.059	0.005	0.100	-0.064
0.61	0.849	0.228	0.490	0.104	0.016	0.017	0.059	0.005	0.100	-0.064
0.62	0.847	0.226	0.487	0.104	0.016	0.018	0.059	0.005	0.100	-0.064
0.63	0.845	0.225	0.484	0.104	0.016	0.018	0.059	0.005	0.100	-0.064
0.64	0.843	0.224	0.481	0.104	0.016	0.018	0.059	0.005	0.100	-0.064
0.65	0.841	0.222	0.478	0.104	0.017	0.019	0.059	0.005	0.100	-0.064
0.66	0.838	0.221	0.475	0.104	0.017	0.019	0.059	0.005	0.100	-0.064
0.67	0.836	0.220	0.472	0.104	0.017	0.020	0.058	0.005	0.100	-0.064
0.68	0.834	0.218	0.470	0.104	0.017	0.020	0.058	0.005	0.100	-0.064
0.69	0.832	0.217	0.467	0.104	0.017	0.020	0.058	0.005	0.100	-0.064
0.70	0.830	0.216	0.464	0.104	0.017	0.020	0.058	0.005	0.100	-0.064

Table 7b

R-SQUARE AND BETA COEFFICIENTS FOR ESTIMATED VALUES OF K										
k	RSQ	HDL	LDL	TRG	CR	P	KA	MPV	CAD	CRP
0.71	0.828	0.215	0.461	0.104	0.018	0.021	0.058	0.006	0.100	-0.064
0.72	0.825	0.213	0.459	0.104	0.018	0.021	0.058	0.006	0.100	-0.064
0.73	0.823	0.212	0.456	0.104	0.018	0.021	0.058	0.006	0.100	-0.064
0.74	0.821	0.211	0.453	0.104	0.018	0.022	0.058	0.006	0.100	-0.064
0.75	0.819	0.210	0.451	0.104	0.018	0.022	0.058	0.006	0.100	-0.064
0.76	0.817	0.209	0.448	0.104	0.018	0.022	0.057	0.006	0.100	-0.064
0.77	0.815	0.207	0.446	0.104	0.019	0.022	0.057	0.006	0.100	-0.064
0.78	0.813	0.206	0.443	0.104	0.019	0.023	0.057	0.006	0.100	-0.064
0.79	0.811	0.205	0.441	0.104	0.019	0.023	0.057	0.006	0.100	-0.064
0.80	0.808	0.204	0.438	0.104	0.019	0.023	0.057	0.006	0.099	-0.064
0.81	0.806	0.203	0.436	0.104	0.019	0.023	0.057	0.006	0.099	-0.064
0.82	0.804	0.202	0.434	0.104	0.019	0.024	0.057	0.006	0.099	-0.064
0.83	0.802	0.201	0.431	0.104	0.019	0.024	0.057	0.006	0.099	-0.064
0.84	0.800	0.199	0.429	0.104	0.019	0.024	0.056	0.006	0.099	-0.064
0.85	0.798	0.198	0.427	0.104	0.020	0.024	0.056	0.006	0.099	-0.064
0.86	0.796	0.197	0.424	0.104	0.020	0.024	0.056	0.006	0.099	-0.064
0.87	0.794	0.196	0.422	0.103	0.020	0.025	0.056	0.006	0.099	-0.064
0.88	0.792	0.195	0.420	0.103	0.020	0.025	0.056	0.006	0.099	-0.063
0.89	0.790	0.194	0.418	0.103	0.020	0.025	0.056	0.006	0.099	-0.063
0.90	0.788	0.193	0.416	0.103	0.020	0.025	0.056	0.007	0.099	-0.063
0.91	0.786	0.192	0.414	0.103	0.020	0.025	0.056	0.007	0.098	-0.063
0.92	0.784	0.191	0.411	0.103	0.020	0.026	0.056	0.007	0.098	-0.063
0.93	0.782	0.190	0.409	0.103	0.020	0.026	0.055	0.007	0.098	-0.063
0.94	0.780	0.189	0.407	0.103	0.021	0.026	0.055	0.007	0.098	-0.063
0.95	0.778	0.188	0.405	0.103	0.021	0.026	0.055	0.007	0.098	-0.063
0.96	0.776	0.187	0.403	0.103	0.021	0.026	0.055	0.007	0.098	-0.063
0.97	0.774	0.186	0.401	0.103	0.021	0.026	0.055	0.007	0.098	-0.063
0.98	0.772	0.185	0.399	0.102	0.021	0.026	0.055	0.007	0.098	-0.063
0.99	0.770	0.184	0.397	0.102	0.021	0.027	0.055	0.007	0.097	-0.063
1.00	0.769	0.183	0.396	0.102	0.021	0.027	0.055	0.007	0.097	-0.063

Table 7c

Figure 1 shows how the beta coefficients for each independent variable are formed as the value of the parameter k increases from 0 to 1. This graph is called ridge trace. We observe that most of the beta coefficients have been stabilized at almost $k = 0.2$. For this value, the regression model is

$$CHOL = 0.288HDL + 0.687LDL + 0.091TRG + 0.000CR - 0.015P + 0.063KA + 0.009MPV + 0.067CAD - 0.052CRP.$$

The R^2 for this model is 94.3%. Here we have to note that the beta coefficients in the above equation are the standardized beta coefficients.

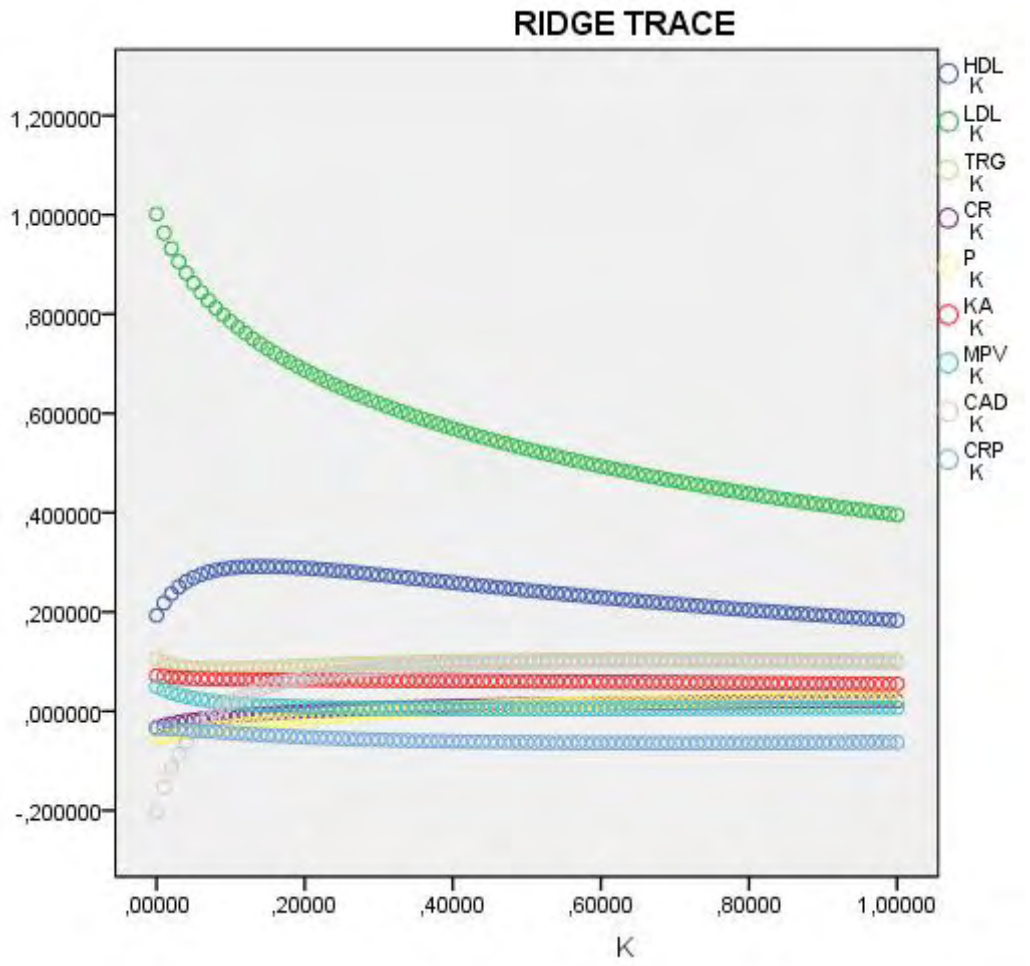

Figure 1

Figure 2 shows how the R^2 value is formed as the value of the parameter k increases from 0 to 1. We observe that as k increases the R^2 value decreases. For $k = 1$ the value of R^2 is 76.9%.

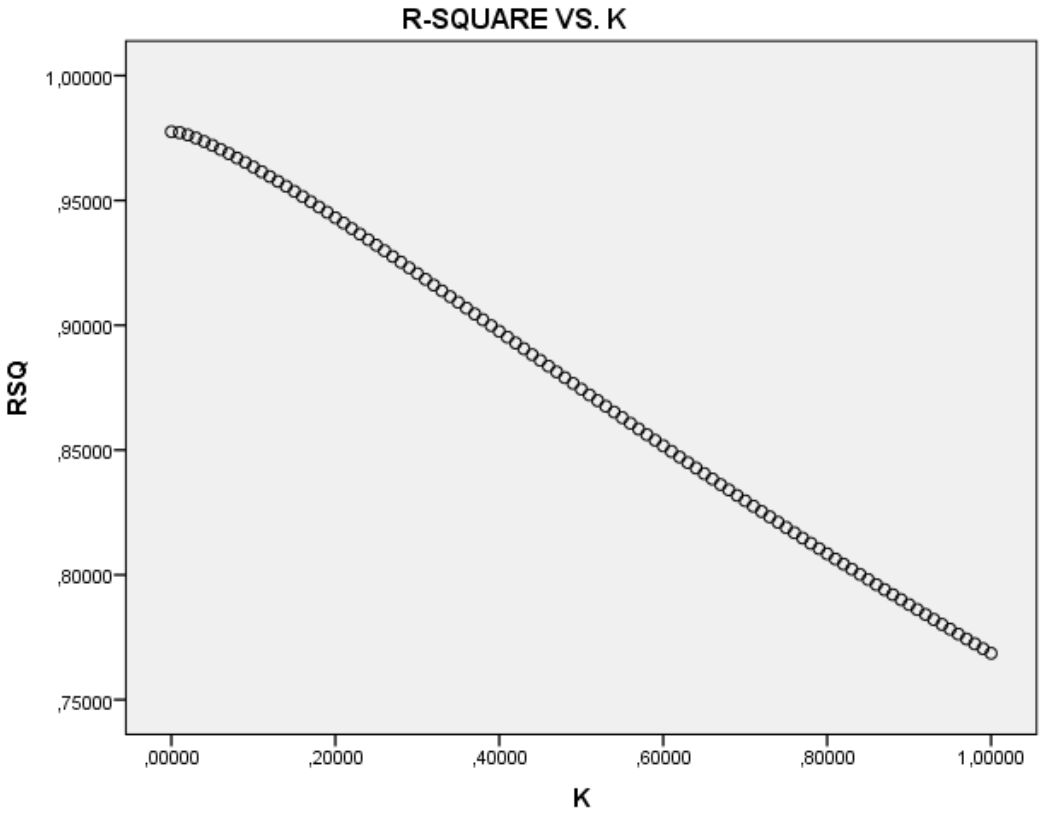

Figure 2

SECTION 5: CONCLUSIONS

In this dissertation, we referred to the multicollinearity problem, the methods of identifying this problem and the ridge regression as a method of handling multicollinearity.

Multicollinearity occurs when two or more independent variables are highly correlated. The main consequence of multicollinearity is that the parameter estimates are less precise.

The identification of multicollinearity is mainly done by examining the correlation matrix, by computing VIF and using eigenvalues of the correlation matrix. Once multicollinearity is detected, it is necessary to modify the regression model.

Remedial measures such as ridge regression and principal component regression help to solve the problem of multicollinearity. Several comparisons of these methods recommend ridge regression (Adnan, 2006; Dorman et al., 2013; Irfan, 2013).

A real data set with biochemical indices for 163 Greek people was used to illustrate how the method is applied.

REFERENCES

Adnan, N., Ahmad, M.H. and Adnan, R. (2006). A Comparative Study on Some Methods for Handling Multicollinearity Problems, *Matematika*, 22 (2), 109–119.

Belsley, D.A. (1991). *Conditioning diagnostics: collinearity and weak data regression*, Wiley.

Belsley, D.A. et al. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*, Wiley.

Booth, G.D. et al. (1994). *Identifying proxy sets in multiple linear regression: an aid to better coeffi cient interpretation* – US Dept of Agriculture, Forest Service.

Chatterjee, S. and Hadi, A. (2006). *Regression Analysis by Examples*, Wiley.

Douglass, D.H. et al. (2003). Test for harmful collinearity among predictor variables used in modeling global temperature, *Climate Research*, 24, 15–18.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67.

Irfan, M., Javed, M. and Raza M.A. (2013). Comparison of Shrinkage Regression Methods for Remedy of Multicollinearity Problem, *Middle-East Journal of Scientific Research*, 14 (4), 570–579.

Johnston, J. (1984). *Econometric methods*, McGraw-Hill.

Maggana, M. *Endothelial dysfunction: the effect of childhood obesity*, Master thesis (in Greek).

McDonald, G.C. (2009). Ridge regression, *WIREs Computational Statistics*, 1, 93–100.

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001). *Introduction to linear regression analysis*, 3rd edition, Wiley, New York.

O'Brien R.M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors, *Quality & Quantity*, 41, 673–690.

Saleh, M.S. (2014). Using Ridge Regression model to solving Multicollinearity problem, *International Journal of Scientific & Engineering Research*, 5 (10), 992–998.

Seber, G.A.F. and Lee, A.J. (2003). *Linear Regression Analysis*, 2nd ed., Wiley.